Creating a social media-based personal emotional lexicon

Ricardo Martins, José João Almeida, Pedro Henriques, Paulo Novais
Algoritmi Centre / Department of Informatics
University of Minho, Braga - Portugal
ricardo.martins@algoritmi.uminho.pt, {jj, prh, pjon}@di.uminho.pt

Keywords: Natural Language Processing, Sentiment Analysis, Machine Learning

Abstract: One of the major problems when using lexicon in sentiment analysis is that they do not cover all possible words in a text and frequently they miss the more expressive to describe the emotions of the text's author efficiently. This problem occurs because people in non-official, on formal channels, communicate using slangs, neologisms, new patterns based on abbreviations (as "aka", "brb" and "asap") and the different meanings, making challenging to analyse texts using a finite subset of a language. This is a problem because some unknown words can completely change the meaning of a sentence, producing misunderstandings. In this paper we present an approach to expand an emotional lexicon for a specific author, producing a customised lexicon which represents how the author "feels" the words. In our experiments, we got an increase of 35.34% and 107.02% in the dictionary size when compared to the original lexicon using two different authors, and identifying different emotions from the same text according to each author's lexicon, i.e. interpreting the text according to the author's "point of view".

1 Introduction

It is common in large countries that people share known words with different meanings. In Portugal, people from Lisbon order a draft beer as "imperial" while in Porto is "fino", however, in Lisbon "fino" can be a polite person and in Porto "imperial" is about everything related to the Portuguese royalty. These words, pronounced by people from different locations express different emotions, and on the other hand, these emotions express the sentiment that the author wanted to transmit when pronounced them. So, it is essential to know what the author wants to transmit, in order to avoid misunderstandings - commonplace when we travel to different countries which speak the same language. If we do not know the author's meaning for used words in the text, we primarily will "decode" the words according to our comprehension of them. So it could be a problem! In sentiment analysis, the same problem occurs when applying emotional lexicon to detect the emotions embedded in the text. Once one or more persons create the emotional lexicon, different emotional interpretation can be applied to the words, and other some other cannot be present in the lexicon. So, detecting emotions using an emotional lexicon as support is like a "pieces of different points of view" instead of author's vision.

In this paper we present an approach for creating a personal emotional lexicon based on social media messages, using Natural Language Processing.

The remainder of the paper is as follows: Section 2 introduces the concept of emotional lexicons, which is the basis of our work. Section 3 discusses related work concerned with the lexicon expansion, while Section 4 describes the approach used to expand and personalise the lexicon. Section 5 presents the results obtained using this approach. The paper ends in Section 6 with the conclusions and future work.

2 Emotion Lexicons

According to Dictionary.com, the term lexicon is defined as:

1. a wordbook or dictionary, especially of Greek, Latin, or Hebrew;
2. the vocabulary of a particular language, field, social class, person, etc.;
3. inventory or record.

3 Related work

There are several works of lexicon expansion for diversified objectives, and some are more relevant for the present paper as they have been used as a source for the idea proposed. In this section, we survey these inspiring works.

The idea of a personal emotional lexicon was inspired by the work of Martins et al. [6], which uses emotional labels to improve the authorship identification. This identification is made using social media posts from chosen personalities and an approach containing lexicon and machine learning approaches, where the usage of a personal lexicon of each author would increase the accuracy of the identification.

The work of Kanayama [3] contributed to the idea of an unsupervised lexicon building method. Although this work handles only with polarities, the process of identifying individual words and share their polarities to other words is an essential issue in our approach.

On your hand, Bravo-Marquez [2] inspired the use of texts from social media and the identification of their particularities, like hashtags, emoticons and neologisms.

4 Lexicon expansion process

The initial point for a lexicon expansion is to collect as many as possible texts from the authors, used to express their thoughts. Since people tend to express themselves differently, using specific words more or less frequently to designate affections, emotions, or even repudiation, the identification of this personal vocabulary will serve as a "fingerprint" of how the author expresses their perceptions.

For this reason, to create this "fingerprint" in our study, we collected comments from Twitter to create a personalised emotional lexicon which corresponds as the way as the author expresses. Twitter was elected as a source of information because differently than other social media because the comments are not restricted a topic, or comments about a post, so people tend to express more freely when there are no constraints in social media.

In our study, it was collected all published tweets from different authors. Due to space limitations, the analysis will present only the data for Donald Trump and Bill Gates.

The process of lexicon expansion, as presented in Figure 1, is composed of different steps, changing the the original and unstructured text in a format able to extract relevant information.

Figure 1: Lexicon expansion process

4.1 Corpus creation

After collecting tweets from the authors, all texts are processed in order to remove unusable information, remaining only the text message written. So, using the Stanford Core NLP [5] toolkit, the Part of Speech (POS) tagger identifies and removes all texts different than nouns, verbs, adverbs and adjectives. The remainder information is stored in a new file, hereafter called corpus.

4.2 Vocabulary creation

For this step, there are two main approaches: one hot encoding vectors - that creates vectors of 0's and 1's to represent the existence of words in a sentence - and word embeddings vectors - that takes in consideration the proximity of known keywords in a sentence. This study applied the word embeddings approach because according to Bayardo [1], when the words in a corpus are distributed in a vector space - as word embeddings are - the similarity can be measured through the *Cosine Distance* between the words. Moreover, according to Kiela [4], has advantages in the identification of similar words, that is the core functionality in the lexicon expansion, that is the idea of our work.

In our comprehension, word embedding is a mapping $\mathrm{V} \rightarrow \mathbb{R}^D : w \rightarrow \overrightarrow{w}$ that maps a word w from a vocabulary V to a word vector $\overrightarrow{w}$ in an embedding space of dimensionality D.

For the word vector's creation, it was applied the Glove [9] using the corpus file of each author sources, resulting in 2 different lists of vectors, representing the authors' vocabulary.

4.3 Similarities

The next step is to analyse the similarity between words. Based on a set of known emotional seed words, the objective is to identify in the corpus the most similar words

related to these seed emotional words. Once identified that words are similar, it is possible to claim that they have the same basic emotions values.

It is possible to find the same similar word related to different seed words (for example, "looking" is similar to seeing and seeming). In this case, "looking" must be disambiguated in order to have the emotions annotated correctly according to his meaning. However, it is not part of this work to handle with disambiguation process actually - it will be handled in future work -, so for the lexicon expansion process, we consider only the similar word containing the higher cosine distance value.

In order to identify these similar words, we used them as emotional seed words the ones contained in EmoLex lexicon [8] while for the similar word's identification, it was created a process to iterate inside the lexicon and a recursive process to iterate into the authors word vectors' in order to detect the similar words. This recursivity allows to identify and associate deep levels of similar words higher than a predefined threshold - in our tests was applied 0.8 as a threshold -, based on the original corpus. If a word has a similarity higher than the threshold, it means that the similar word shares the same basic emotions values with the lexicon word. All identified similar words and basic emotions and their lexicon emotional words and their basic emotions are stored in order to input the next step, hereafter called "Similarities".

4.4 Synonyms

The next step detects the synonyms of each word in "Similarities". To reach this objective, all words in "Similarities" were analysed in the Wordnet [7] in order to identify all synonyms for the word. An important detail in this step is the attention with pre-existent words. Once the idea is detecting the emotions related to how the author expresses in a text, the most critical information is the words identified as similar. So, in the case of words identified as similar and synonyms, the synonym is discarded.

Like the previous step, after the synonyms identification was created a recursive process to iterate inside the synonyms and the author's word vectors' in order to detect the similar words between "Similarities" and synonyms. All identified similar words and their basic emotions and the synonym of the emotional words and their basic emotions are stored, resulting in the personal expanded lexicon.

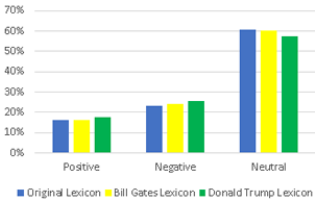

Figure 2: Lexicon polarities

5 Results

Once performed all process described in the previous sections, it was generated different lexicons, based on the author's vocabulary.

Using the original seed lexicon as a parameter, it is possible to compare the amount of information added for each author. Table 1 shows author's lexicon words increasing for each emotion and their respective rate when compared to the original lexicon.

Based on these pieces of information, according to Figure 3, Bill Gates lexicon has almost the same proportion of the basic emotions when compared to the original lexicon (2 of 8). On the other hand, Donald Trump lexicon presents proportional differences in 6 of 8 basic emotions when compared to the original lexicon. Also, as presented in Figure 2, Donald Trump lexicon has proportionally more positive and negative words, and fewer neutral words, when compared to original and Bill Gates lexicon, demonstrating that the first author is blunter than the other author, raising more emotions in their speeches.

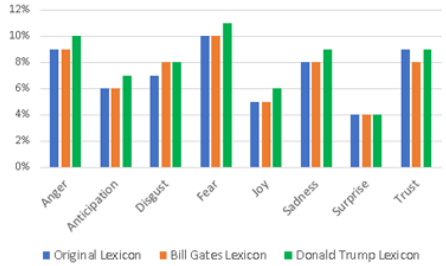

Figure 3: Proportions

Furthermore, different personal lexicon can represent an opportunity of interpreting texts as the lexicon author could interpret them. In order to perform this analysis, we choose two different texts from the authors: the Bill Gates' World Economic Forum speech (2008) and Donald Trump's United Nations speech (2017).

Each text was analysed using three different lexicons: original lexicon, Bill Gates lexicon and Donald Trump lexicon, in order to identify the existent words in the texts and their related emotions existent in each lexicon. Tables 2 and 3 present the proportion of each

Table 1: Lexicon comparison

		Total Words	Positive	Negative	Neutral	Anger	Anticipation	Disgust	Fear	Joy	Sadness	Surprise	Trust
	Original Lexicon	14182	2312	3324	8546	1247	839	1058	1476	689	1191	534	1231
Donald Trump Lexicon	**Words**	19194	3098	4638	11458	1679	1141	1549	1923	937	1510	750	1571
	Increasing Rate	35,34%	34,00%	39,53%	34,07%	34,64%	36,00%	46,41%	30,28%	35,99%	26,78%	40,45%	27,62%
Bill Gates Lexicon	**Words**	29360	5186	7461	16713	2901	1937	2482	3201	1616	2628	1237	2740
	Increasing Rate	107,02%	124,31%	124,46%	95,57%	132,64%	130,87%	134,59%	116,87%	134,54%	120,65%	131,65%	122,58%

emotion in Bill Gates and Donald Trump speeches respectively.

An impressive result is that when we use a lexicon from another author to analyse the text, the sentiments *Anger, Fear* and *Sadness* are lower than the values obtained when using the author's lexicon. On the other hand, the *Joy* sentiment is higher in texts from authors different than analyses. In this case, it can be interpreted as the author can be more "complacent" with others and more "stern" with himself.

Table 2: Bill Gates' speech analysis

Emotion	Original	Donald Trump	Bill Gates
Anger	5.57%	4.09%	11.47%
Anticipation	18.38%	22.80%	13.09%
Disgust	4.18%	5.59%	6.48%
Fear	10.58%	7.31%	10.39%
Joy	17.55%	18.92%	12.28%
Sadness	6.41%	6.24%	12.42%
Surprise	7.52%	7.31%	5.94%
Trust	29.81%	27.74%	27.94%

Table 3: Donald Trump speech analysis

Emotion	Original	Donald Trump	Bill Gates
Anger	10.36%	11.44%	8.51%
Anticipation	15.00%	11.90%	12.71%
Disgust	6.18%	7.28%	7.01%
Fear	15.82%	13.49%	12.71%
Joy	12.82%	11.57%	17.52%
Sadness	9.18%	9.19%	7.71%
Surprise	4.73%	5.95%	5.31%
Trust	25.91%	29.17%	28.53%

6 Conclusion

This work presents an approach to create a personal lexicon based on the expansion of a seed lexicon through social media texts. This personal lexicon contains the vocabulary used for each author and the emotions associated with each word, according to what the author wanted to transmit. This solution decreases the problem of misunderstandings when interpreting texts because it helps to know the emotions that the author wanted to transmit in their text. Once the interpretation of emotions is mainly personal, knowing the word's meaning according to each author helps to interpret the text according to the author's point of view. Moreover, new expressions - even hashtags - and local expressions raise quickly, and "translating" its emotional meaning takes time when compared to traditional emotional lexicons.

The word's increasing of 35.34% and 107.02% of each personal lexicon, when compared to the original lexicon, shows that this solution can be used in sentiment analysis processes because it increases the possibility of text interpretation. Regarding Natural Language Processing, it contributes to provide a customised service to the end user, enabling to avoid misunderstandings when interpreting texts, analysing sentiments in an individual scale and increasing the level of accuracy about recommendations, based on their characteristics and personality.

As future work, it is planned to handle the lexicon expansion for different domains, which allows understanding how the sentiments can change according to the context of a conversation, as well as the intensity for each basic emotion in the domains.

Acknowledgements

This work has been supported by COMPETE: POCI-01-0145-FEDER-0070 43 and FCT – Fundação para a Ciência e Tecnologia within the Project Scope UID/CEC/ 00319/2013.

REFERENCES

[1] Roberto J Bayardo, Yiming Ma, and Ramakrishnan Srikant. Scaling up all pairs similarity search. In *Proceedings of the 16th international conference on World Wide Web*, pages 131–140. ACM, 2007.

[2] Felipe Bravo-Marquez, Eibe Frank, and Bernhard Pfahringer. Positive, negative, or neutral: Learning an expanded opinion lexicon from emoticon-annotated tweets. In *IJCAI*, pages 1229–1235, 2015.

[3] Hiroshi Kanayama and Tetsuya Nasukawa. Fully automatic lexicon expansion for domain-oriented sentiment analysis. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 355–363. Association for Computational Linguistics, 2006.

[4] Douwe Kiela, Felix Hill, and Stephen Clark. Specializing word embeddings for similarity or relatedness. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2044–2048, 2015.

[5] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60, 2014.

[6] Ricardo Martins, José Almeida, Pedro Henriques, and Paulo Novais. Increasing authorship identification through emotional analysis. In *World Conference on Information Systems and Technologies*, pages 763–772. Springer, 2018.

[7] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[8] Saif Mohammad and Peter D. Turney. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465, 2013.

[9] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.